Summarization on Text Mining

A.Bharath

Assistant Professor, Department of Computer Science Engineering,
Guru Nanak Institute of Technology, Ibrahimpatnam, Ranga Reddy, Telangana, India

ABSTRACT: -

In this scenario, we focus on multimodal news aggregation retrieval and fusion. In particular, we present preliminary experiments aimed at automatically suggesting keywords to news and news aggregations. The proposed solution is based on the adoption of extraction-based text summarization techniques. Experiments are aimed at comparing the selected text summarization techniques with respect to a simple technique based on part-ofspeech tagging. Results show that the proposed solution performs better than the baseline solution in terms of precision, recall, and F1. For example, human sentiments can be positive, negative. Now a Days we highly consider opinions of friends, domain experts for decision making in day today's life. Natural language techniques are applied to extract emotions from unstructured data. In marketing and advertising domains Opinion Mining being larger domain. The advertiser required to the analyze performance/ ads status that person posted on site. Star rating based on mechanism may go fraud, automatic robots or responders. So, the present system required to analyze applying NLP & comments. Fraud comments could indifferent through applying irrelevant comment elimination mechanism suggested in the paper. In that paper the role and importance of opinions on public are discussed especially. Various techniques that proposed and emerged to discuss about the opinions are mentioned in details.

INTRODUCTION

Document summarization refers to the task of creating document surrogates that are smaller in size but retain various characteristics of the original document. To automate the process of abstracting, researchers generally rely on a two phase process. First, key textual elements, e.g., keywords, clauses, sentences, or paragraphs are extracted from text using linguistic and statistical analyses. In the second step, the extracted text may be used as a summary. Such summaries are referred to as „extracts‟. Alternatively, textual elements can be used to generate new text, similar to the human authored abstract. Summarization of Hindi documents contains historical information is also plays as important role for students and teachers who want to read a large number of documents related to history. Summarization system helps them to read and learn the shorter version of overall complete document Summarization system helps them to read and learn the shorter version of overall complete document.

Automatic Text Summarization is an important and challenging area of Natural Language Processing. The task of a text summarizer is to produce a synopsis of any document or a set of documents submitted to it Analysis of Text-Documents has been an active area of research for the past few years. It involves extensive use of Natural Language Processing techniques for analyzing semantic structures of the text.

Semantic analysis of a document means to analyze the meaning or transitions in meaning of the sentences or of different clauses and the relation among them. There are a number of approaches to semantic analysis. Semantic analysis can be done at the sentence level, the paragraph level, or even at the text level on different languages.

SUMMARIZATION CAN BE OF TWO TYPES:

Extractive and Abstractive. In our proposed system, we have chosen extractive summarization for the study purpose. What characteristics a sentence should possess to grab the position in the summary, is the core question to be answered. These characteristics are called as features and extraction of these features calculates the overall score a sentence would weigh. In our system, we have suggested six statistical and two linguistic features to be extracted. We are proposing two machine learning techniques Genetic Algorithm (GA) and Artificial Neural Network (ANN) for the sentence extraction and ranking. It is then followed by the comparative study of both the algorithms.

We have considered Hindi as a language of Study. It is written in the Devanagari script which has largest alphabet set. Hindi is an official language of India. It the native language of most people living in Delhi, Chhattisgarh, Himachal Pradesh, Chandigarh, Bihar, Jharkhand, Madhya Pradesh, Haryana, and Rajasthan. So for people who do not know English but want to read articles on the Internet, automatic summarization would play lion‟s role in it. While performing related search, it is observed that a lot of work has been done on English language as ample amount of resources.

Text summarization is the process of distilling the most important information from the set of sources to produce an abridged version [1].

B. Types of Text Summarization

Text summarization can be performed in two different approaches: extraction and abstraction.

1)Extraction: This approach is to construct the summary by producing the most important sentences verbatim out of the original document and is mainly concerned with what the summary content should be.

2) Abstraction: The abstraction approach is to form summary by paraphrasing sections of the original document putting strong emphasis on the form, aiming to produce an important material in a new way.

C. Types of Extraction Method Extraction method is further classified as: Statistical, Linguistic and Hybrid approach.

1) Statistical Method: Text summarization based on this approach relies on the statistical distribution of certain features and it is done without understanding whole document. Models rank the sentences of the original text to appear in the summary in the order of importance. We are using average TF-ISF, title Word, sentence length, sentence feature, thematic word and numerical data as statistical features in our proposal.

2) Linguistic Method: In this, method needs to be aware of and know deeply the linguistic knowledge, so that the computer will be able to analyze the sentences and then decide which sentence to be selected. We are using proper noun feature and sentence to sentence similarity as linguistic features in our proposal.

3) Hybrid Method: It optimizes best of both the previous method for meaningful and short summary.

Various methods have been proposed to achieve extractive summarization. Most of them are based on scoring of the sentences. Dr.Latesh Malik, et. al.[1], Discussed single document summarization using extraction method for Hindi text, which uses statistical and linguistic features. It uses Hindi Wordnet to tag appropriate POS of word for checking SOV of the sentences which uses sixstatistical and two linguistic features. It also uses genetic algorithm to optimize the summary generated based on the text feature terms with less redundancy. Ibrahim F. Moawad, et. al.[2], Described a noval approach is presented to create an abstractive summary for a single document using a rich semantic graph reducing technique. The approach summaries the input document by creating a semantic graph called Rich Semantic Graph for the original document, reducing

d graph but in English. Sachin Agarwal, et. al.[3], Proposed the algorithm for anaphora resolution has been tested extensively. The accuracy of anaphora resolution is 96% for simple sentence not for original document and complex sentences; the accuracy is of the order of 80%. This method works by searching sentences in the text that are semantically related thorough anaphors, analyzing their semantic s and exploiting the latter for s resolving respective anaphors. Ng Choon-Ching, et. al.[4], Proposed an existing need for text summarizers that small devices like PDA has emerged the development of text summarization of web pages. Authors have identified problems for text summarization in several areas such as dynamic content of web pages, diverse summary definition of text, and different benchmark of evaluation measurements.

Besides, authors also found advantages of certain methods that increased the accuracy of web page classification. In the future work, author plan to investigate machine learning techniques to incorporate additional features for the improvement of text summarization quality. The additional features authors are currently considering include linguistic features such as discourse structure, lexical chains, semantic features such as name entities, time, location information etc Visual Gupta, et. al.[5], Describe the Punjabi text extractive system which consist of two phases

1) Pre Processing

2) Processing.

In this paper term pre processing is defined as the phase which identify the word boundary, sentence boundary, Punjabi stop words elimination etc. and the processing phase sentence features are calculated and a weight is assigned to each sentence on the reference of which unwanted sentences are eliminated from the input text. It is described that the author tested the proposed system over fifty Punjabi news documents (with 6185 sentences and 72689 words) from Punjabi Ajit news paper and fifty Punjabi stories (with 17538 sentences and178400 words). Accuracy of the system is varies from 81% to 92 %.

Niladri Chatterjee, et. al.[6], Described summarization technique for text document exploiting the semantic similarity between sentences to remove the redundancy from the text. It uses Random Indexing for compute the semantic similarity scores of sentences and graph-based ranking algorithms have been employed to produce an extract of the given text.

The important is that the problem of high dimensionality of the semantic space corresponding text should be tackled by random indexing which is less expensive in computations and memory consumption and it included a training algorithm using Random Indexing which has to construct the Word space on complied text document so that resolve the ambiguities such as more efficiency.

M. C. Padma, et. al.[7], In a multi-script multi-lingual environment, a document may contain text lines in more than one script/language forms. It is necessary to identify different script regions of the document in order to feed the document to the OCRs of individual language. With this context, this paper proposes to develop a homothetic algorithmic model to identify and separate text lines Telugu, Hindi and English scripts from a printed multilingual document.

The proposed method uses the distinct features of the target script and searches for the text lines that possess the anticipated features. Experimentation conducted involved 1500 text lines for learning and 900 text lines for testing. The performance has turned out to be 98.5%.

They have no semantic as such and do not aggregate relevant information to the task. Also they make the text look heavier and are insignificant. Hence should be eliminated.

4) Stemming : In Stemming process, the suffixes are ignored and removed from words to get the common origin. It recognizes words with common meaning and form as being identical. Syntactically similar words, such as plurals, verbal variations, etc. are considered similar. e.g. walk, walking and walked are counted as same and derived from a stem word walk. B. Processing Step In processing step, we decide and calculate the features that affect the relevance of sentences and then weights are assigned to these features using weight learning method. Higher ranked sentences are extracted for summary.

Feature Extraction: Real analysis of the document for summarization begins in this phase. Every sentence is represented by the feature terms vector and has a score based on the weight of feature terms. This score is used for sentence ranking. Feature term values range between 0 to1. Six statistical and two linguistic features are used as follows:

1) Average TF-ISF (Term Frequency Inverse Sentence Frequency): TF-ISF stands for term frequency-inverse document frequency and the tf-isf weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the sentence (TF) but is offset by the frequency of the word in the corpus (ISF). We should look at the distribution of the word across the complete document instead of making only a local comparison

Sentence Length: The short sentences such as datelines and author names are not expected to belong to the summary. In the same way, too long sentences may contain a lot of redundant data and hence are unlikely to be included in the summary. So, we eliminate the sentences which are too short or too long. This feature computation uses minimum and maximum length threshold values. Consider L = Length of Sentence MinL = Minimum Length of Sentence (= 5 in our experiment) MaxL = Maximum Length of Sentence (=15 in our experiment) Min Θ = Minimum Angle (0) and Max Θ = Maximum Angle (180).

ANALYSIS

Summarization techniques can be divided in two groups [16]:

those that extract information from the source documents (extraction-based approaches) and those that abstract from the source documents (abstraction-based approaches). The former impose the constraint that a summary uses only components extracted from the source document. These approaches put strong emphasis on the form, aiming to produce a grammatical summary, which usually requires advanced language generation techniques. The latter latter relax the constraints on how the summary is created. These approaches are mainly concerned with what the summary content should be, usually relying solely on extraction of sentences. Although potentially more powerful, abstraction-based approaches have been far less popular than their extraction-based counterparts, mainly because generating the latter is easier. While focusing on information retrieval, one can also consider topic driven summarization, which assumes that the summary content depends on the preferences of the user and can be assessed via a query, making the final summary focused on a particular topic. Since in this paper we are interested in extracting suitable keywords, we exclusively focus on extraction-based methods. An extraction-based summary consists of a subset of words from the original document and its bag of words (BoW) representation can be created by selectively removing a number of features from the original term set. Typically, an extractionbased summary whose length is only 10-15% of the original is likely to lead to a significant feature reduction as well. Many studies suggest that even simple summaries are quite effective in carrying over the relevant information about a document. From a text categorization perspective, their advantage over specialized feature selection methods lies in their reliance on a single document (the one that is being summarized) without computing the statistics for all documents sharing the same category

CONCLUSION

When information on the Internet began growing exponentially in the 1990s, computer scientists recognized the need to search for and retrieve information. To help users extract information and generate summaries, they worked with linguists to formalize the features of summaries and incorporate them in their programs. In the literature, a frequently cited definition is given in Radev (2002; cited in Das and Martins, 2007): summaries can be of one document or of several documents; they should be short; and they should preserve important information. As Das and Martins (2007) point out, "a more elaborate definition for the task would result in disagreement within the community" (p.1), which we see in moderation meetings for examinations. This definition covers user inputs—the number of documents and the length of the summary. However, it is the final feature, namely, preserve important information, which is the central concern for the student, the writer and researchers working on text summarization techniques. What features of a text help identify 'importance'? Early work in text summarization identified three features of a summary that still hold good. Note that Features 2 and 3 draw on the concept of text structure.

Reference

1. *Armano, G., Giuliani, A., Vargiu, E.: Experimenting text summarization techniques for contextual advertising. In: IIR'11: Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop (2011)*
2. *Armano, G., Giuliani, A., Vargiu, E.: Studying the impact of text summarization on contextual advertising. In: 8th International Workshop on Text-based Information Retrieval (2011)*
3. *Baxendale, P.: Machine-made index for technical literature - an experiment. IBM Journal of Research and Development 2, 354–361 (1958) 4.*
4. *Bertini, M., Del Bimbo, A., Torniai, C.: Enhanced ontologies for video annotation and retrieval. In: Proc. of the 7th ACM SIGMM international workshop on Multimedia information retrieval, pp. 89–96 (2005)*
5. *Bertini, M., Del Bimbo, A., Torniai, C.: Automatic annotation and semantic retrieval of video sequences using multimedia ontologies. In: Proc. of the 14th annual ACM international conference on Multimedia, pp. 679–682 (2006)*
6. *Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic*

publications by sentence selection. Information Processing Management 31, 675–685 (1995)

7. *Das, D., Martins, A.F.: A survey on automatic text summarization. Tech. Rep. Literature Survey for the Language and Statistics II course at CMU (2007)*

8. Deschacht, K., Moens, M.F.: Finding the Best Picture: Cross-Media Retrieval of Content. In: Proc. of ECIR 2008, pp. 539–546 (2008)